Machine Learning 1

Lecture 11.6 - Kernel Methods
Support Vector Machines - Soft Margin Classifier

Erik Bekkers

(Bishop 7.1.1)

Slide credits: Rianne van den Berg

Image credit: Kirillm | Getty Images

Maximum Margin Classifiers

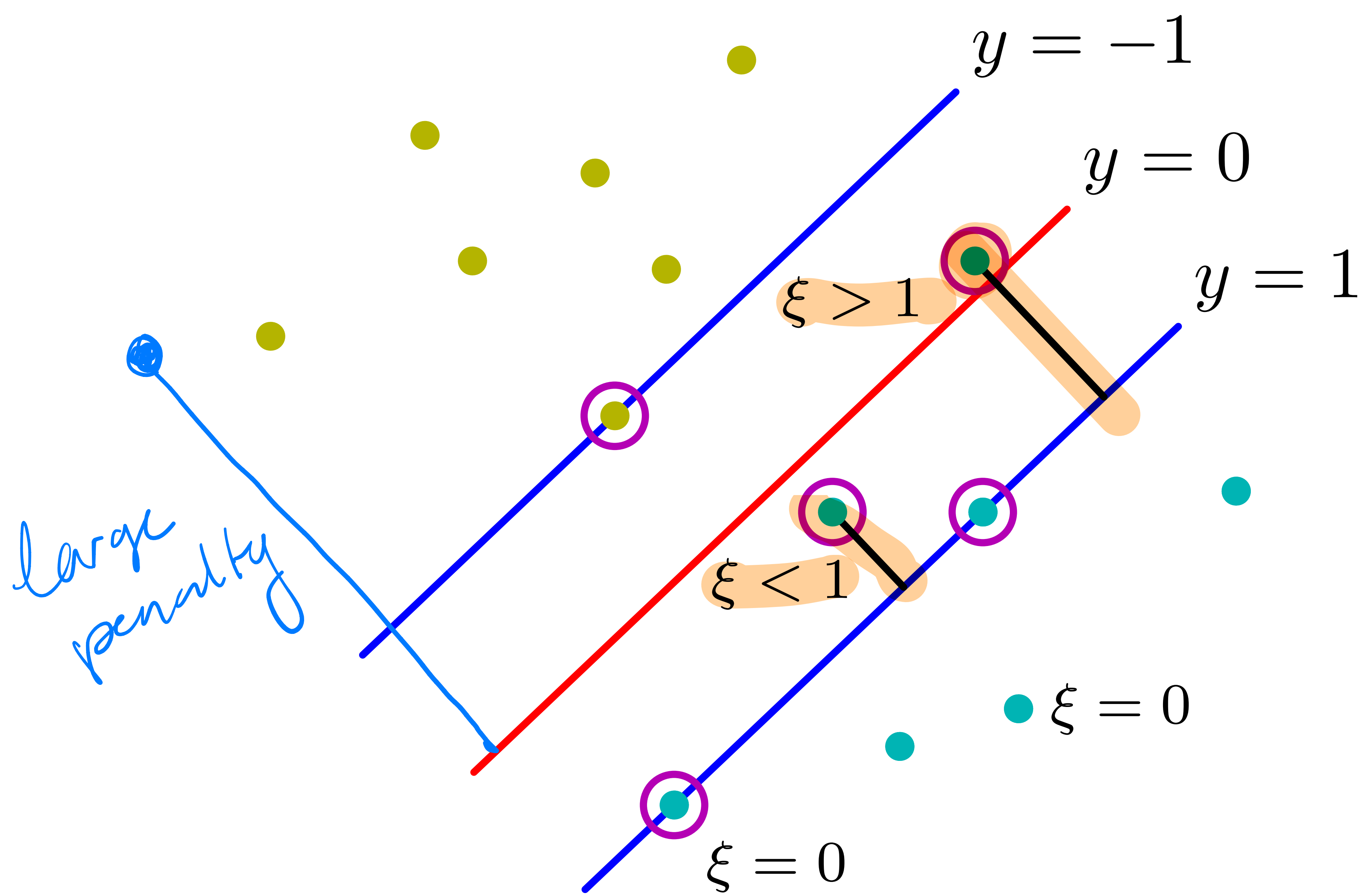
- So far we have assumed the data points are perfectly separable with a linear decision boundary, or with a nonlinear decision boundary by using a nonlinear kernel.

- Sometimes the class conditional distributions have overlap!

- We need to modify the Maximum Margin classifier to allow for some training points to be misclassified.

- Datapoints are allowed to be on the “wrong” side of the margin boundary, but they have to pay a penalty proportional to the distance to the margin boundary.

Soft Margin Classifiers



- Allows datapoints to lie on the wrong side of the margin boundary.
- Those datapoints pay a penalty proportional to the distance to the margin boundary.

Maximum Margin Classifiers: Soft Margins!

- Introduce slack variables: $\xi_n \geq 0$ for $n = 1, \ldots, N$
- If on the correct side of the margin: $\xi_n = 0$
- If on the wrong side of the margin: $\xi_n = |t_n - y(\mathbf{x}_n)|$

±1

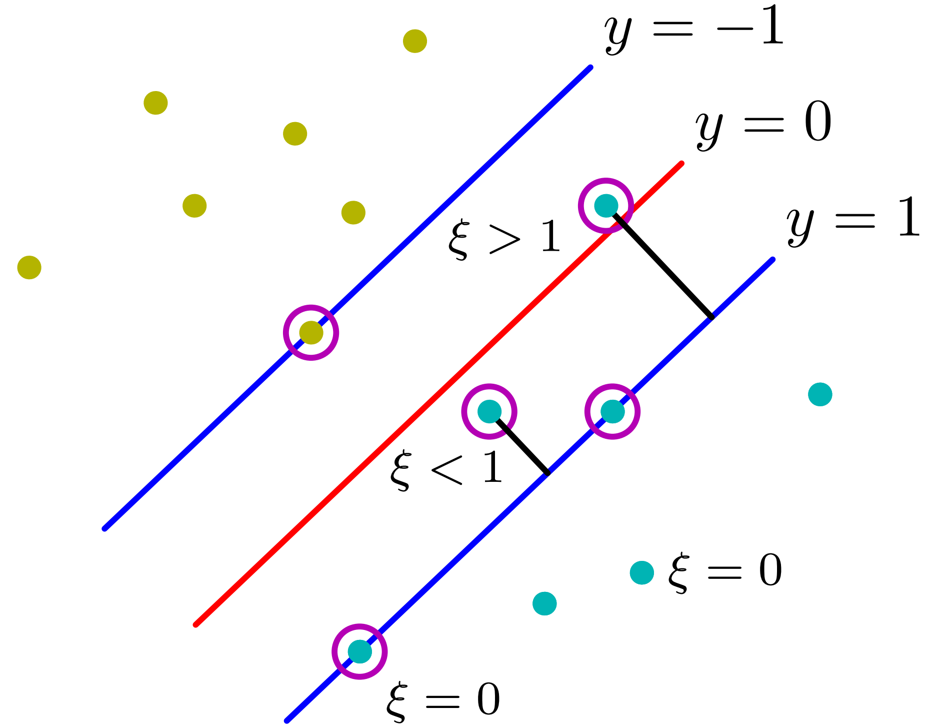
prop. to distance to decision boundary

- Previously: hard constraints/hard margin

$$t_n y(\mathbf{x}_n) \geq 1, \qquad n = 1, \ldots, N$$

- Now: Soft constraint/soft margin

$$t_n y(\mathbf{x}_n) \geq 1 - \xi_n, \; n = 1, \ldots, N$$



Maximum Margin Classifiers: Soft Margins!

max $\frac{1}{\|w\|}$

- Goal: maximize margin, give a penalty to points that lie on the wrong side of the boundary!

- We minimize $\underset{\mathbf{w},b,\xi_n}{\arg\min} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N} \xi_n$

  subject to constraints $\quad t_n y(\mathbf{x}_n) \geq 1 - \xi_n\,,\quad$ for $\quad n = 1,\ldots,N$

  $\xi_n \geq 0\,,\quad$ for $\quad n = 1,\ldots,N$

- Corresponding Lagrangian:

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \mathbf{a}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N} a_n\{t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 + \xi_n\} - \sum_{n=1}^{N}\mu_n\xi_n$$

- Lagrange multipliers

$$a_n \geq 0, \quad \mu_n \geq 0$$

1. Lagrangian
2. KKT
3. Solve primal variables
4. Obtain dual Lagrangian
5. Solve for duals

Maximum Margin Classifiers: Soft Margins!

- Lagrangian function

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \mathbf{a}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N} a_n\{t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 + \xi_n\} - \sum_{n=1}^{N}\mu_n\xi_n$$

- Lagrange multipliers. $a_n \geq 0, \quad \mu_n \geq 0$

- KKT conditions:

$$\begin{aligned} a_n &\geq 0 & \mu_n &\geq 0 \\ t_n y(\mathbf{x}_n) - 1 + \xi_n &\geq 0 & \xi_n &\geq 0 \\ a_n\{t_n y(\mathbf{x}_n) - 1 + \xi_n\} &= 0 & \mu_n\xi_n &= 0 \end{aligned}$$

3N 3N

- How many KKT conditions?

6N

Maximum Margin Classifiers: Soft Margins!

- Lagrangian function

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \mathbf{a}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N} \xi_n - \sum_{n=1}^{N} a_n\{t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 + \xi_n\} - \sum_{n=1}^{N} \mu_n\xi_n$$

- Minimize L w.r.t. primal variables $\mathbf{w}, b, \xi_n$ and use the KKT conditions to eliminate $\mathbf{w}, b, \xi_n$ from Lagrangian to obtain dual formulation!

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w}^T - \sum_{n=1}^{N} a_n t_n \mathbf{x}_n^T = 0 \quad \rightarrow \quad \boxed{\mathbf{w} = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n}$$

$$\frac{\partial L}{\partial b} = -\sum_{n=1}^{N} a_n t_n = 0 \quad \rightarrow \quad \boxed{\sum_{n=1}^{N} a_n t_n = 0}$$

same as before!

$$\frac{\partial L}{\partial \xi_n} = C - a_n - \mu_n = 0 \quad \rightarrow \quad \boxed{a_n = C - \mu_n}$$

new constraint!

- Use this to eliminate $\mathbf{w}, b, \xi_n$, dual Lagrangian:

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m \mathbf{x}_n^T \mathbf{x}_m$$

Maximum Margin Classifiers: Soft Margins!

- Minimization of primal variables gave these conditions:

$$\boxed{\mathbf{w} = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n}\,, \qquad \boxed{\sum_{n=1}^{N} a_n t_n = 0}\,, \qquad \boxed{a_n = C - \mu_n} \quad (*)$$

- KKT conditions:

$$a_n \geq 0 \quad (**) \qquad \mu_n \geq 0$$

$$t_n y(\mathbf{x}_n) - 1 + \xi_n \geq 0 \qquad \xi_n \geq 0$$

$$a_n\{t_n y(\mathbf{x}_n) - 1 + \xi_n\} = 0 \qquad \mu_n \xi_n = 0$$

- Dual lagrangian:

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m \mathbf{x}_n^T \mathbf{x}_m$$

- Remaining constraints:

(**) (*)

$$0 \leq a_n \leq C$$ Box constraints

$$\sum_{n=1}^{N} a_n t_n = 0$$

Maximum Margin Classifiers: Soft Margins!

- Dual problem: Maximize w.r.t an

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m \mathbf{x}_n^T \mathbf{x}_m$$

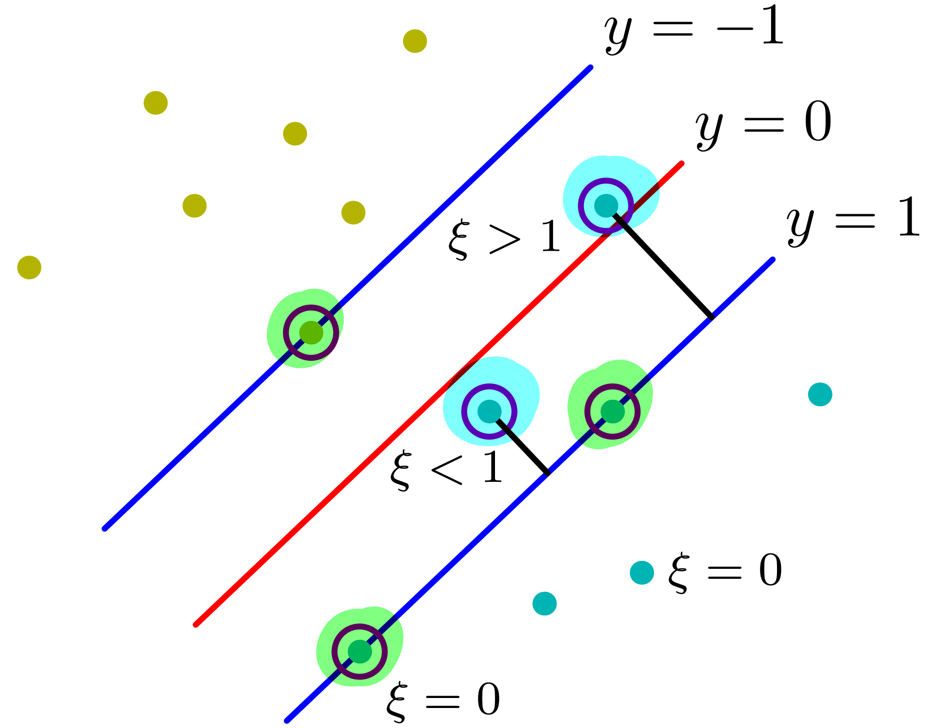
subject to $\quad 0 \leq a_n \leq C\,, \qquad \sum_{n=1}^{N} a_n t_n = 0$

- Kernel trick:

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

- Prediction

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}_n, \mathbf{x}) + b$$



Maximum Margin Classifiers: Soft Margins!

- Prediction: $y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}_n, \mathbf{x}) + b$ (with $0 \leq a_n \leq C$ and $\sum_{n=1}^{N} a_n t_n = 0$)

3 types of SV's

1. $a_n = 0$
2. $0 < a_n < C$
3. $a_n = C$

- Remember $a_n \geq 0$ $\quad \mu_n \geq 0$

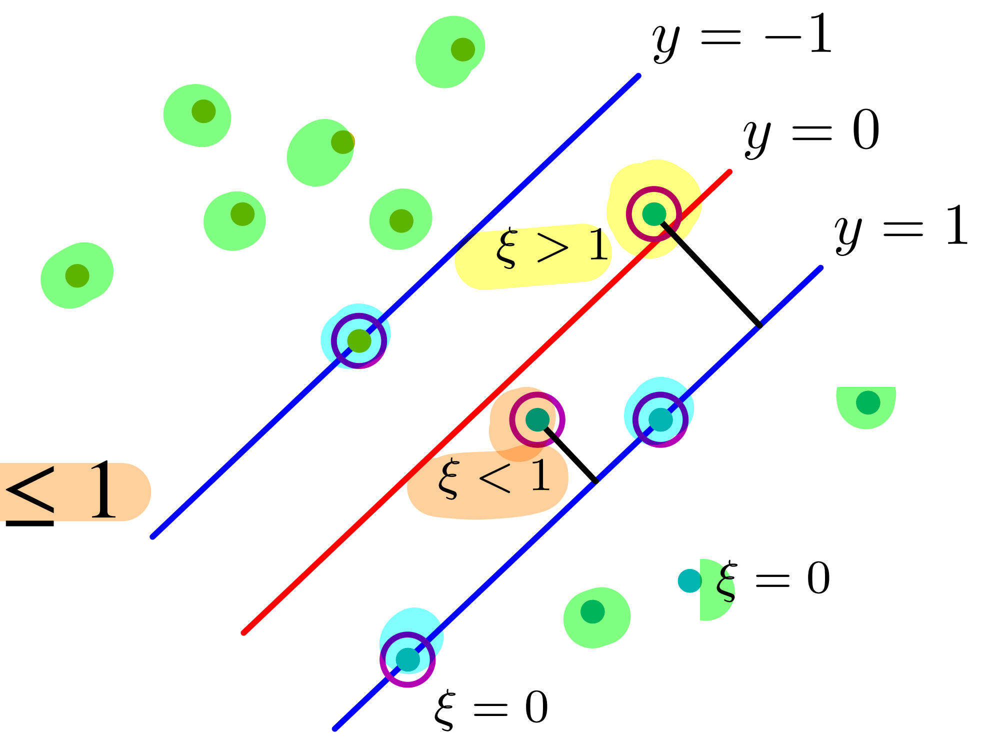
$$t_n y(\mathbf{x}_n) - 1 + \xi_n \geq 0 \qquad \xi_n \geq 0$$

$$a_n\{t_n y(\mathbf{x}_n) - 1 + \xi_n\} = 0 \qquad \mu_n \xi_n = 0$$

- **Support vectors** (If $a_n > 0$ then $t_n y(\mathbf{x}_n) = 1 - \xi_n$):

$(a_n = C - \mu_n)$

  - If also $a_n < C$ then $\mu_n > 0$ so $\xi_n = 0$:
    - Points on margin
  - If $a_n = C$ then $\mu_n = 0$ so $\xi_n \geq 0$:
    - Correctly classified but within margin: $\xi_n \leq 1$
    - Misclassified $\xi_n > 1$



Maximum Margin Classifiers: Soft Margins!

- Goal: maximize margin, give penalty to points that lie on the wrong side of the boundary!

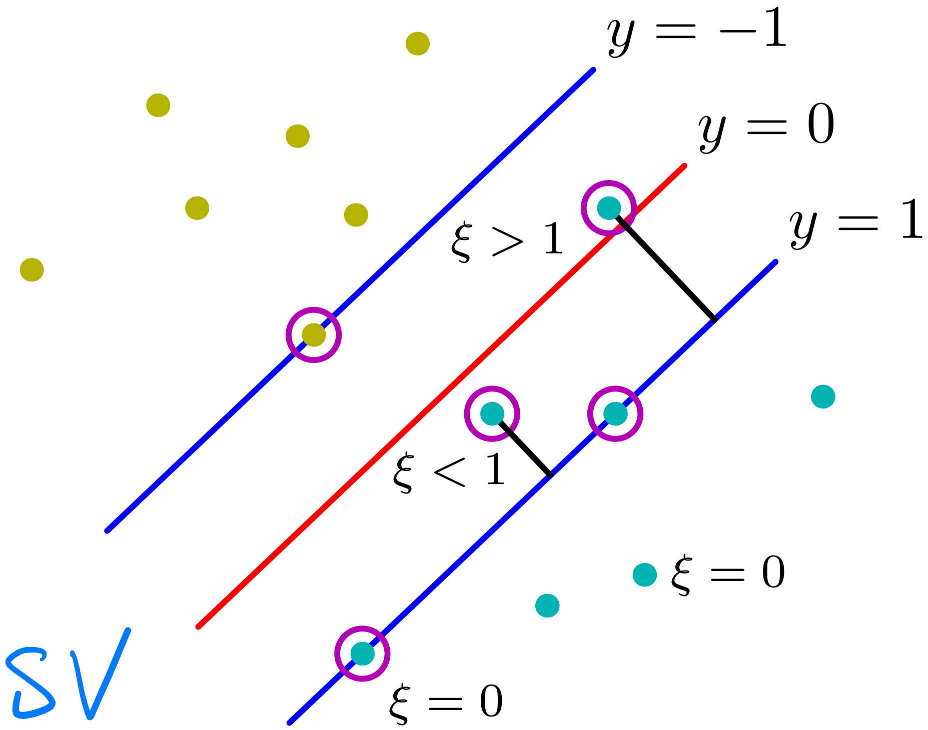
- We minimize $\arg\min_{\mathbf{w},b,\xi_n} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N} \xi_n$

  subject to $t_n y(\mathbf{x}_n) \geq 1 - \xi_n$, for $n = 1,\ldots,N$

  $\xi_n \geq 0$, for $n = 1,\ldots,N$

- What happens in the limit: $C \to \infty$

  - Hard margin Classifier



- What happens when $C \to 0$

  Possibly
  - Infinite margin

  every point becomes a SV